Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 12:43:41
REFERENCE NO: 237

Author Names & Affiliations

- Adam Bolton - National Optical Astronomy Observatory
- Steven Berukoff - National Solar Observatory
- Frossie Economou - Large Synoptic Survey Telescope
- David Liska - Space Telescope Science Institute
- Arfon Smith - Space Telescope Science Institute
- James Turner - Gemini Observatory

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

Mathematics & Physical Sciences, Astronomical Sciences, Observational Astronomy, Optical & Infrared Astronomy

Title of Submission

Cyberinfrastructure for Data-Intensive Survey Astronomy

Abstract (maximum ~200 words).

The last two decades have seen the rise of a new mode of astronomical research, with ambitious observational surveys producing large and homogeneous data sets that support diverse scientific research activities. Going forward, broad-based research throughout the US astronomical community will rely as much on new modes of access to these data sets—and to the scientific opportunities that they provide—as on traditional modes of access to PI-driven observing time with telescopes. The scientific success of data-intensive astronomy depends critically upon the availability and effective use of major cyberinfrastructure resources for storage, processing, analysis, and distribution of petascale data sets. This submission describes the associated challenges and requirements from the perspective of major centers and facilities managed by the Association of Universities for Research in Astronomy (AURA): the NSF-funded National Optical Astronomy Observatory (NOAO), Large Synoptic Survey Telescope (LSST) Project, Gemini Observatory, and National Solar Observatory (NSO), as well as the NASA-funded Space Telescope Science Institute (STScI).

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

The classical picture of research in astronomy features the lone astronomer at the eyepiece of the telescope. While the human eye has been supplanted as the recording device of choice—first by the photographic plate, then in turn by the digital electronic detector—direct

access by individuals and small teams to the resource of observing time with a telescope has remained the primary measure of scientific advantage in optical and infrared astronomy until very recently.

The last two decades have seen the emergence of a powerful new mode of astronomical research, as large-scale scientific questions have motivated ambitious astronomical survey projects that deliver huge data sets. These data sets enable a wide range of scientific research projects, many of which are conceived and executed by individuals and collaborations independent of the original survey team. In this new mode of data-intensive astronomy, access to data sets and analysis capabilities displaces access to observing time as the key research-enabling factor.

In ground-based optical astronomy, the Sloan Digital Sky Survey (SDSS) provides the canonical reference for the big-surveys/big-data mode of science. In operation since 2000, and now in its fourth phase, the SDSS has used a dedicated 2.5-meter telescope at Apache Point Observatory to carry out multiple surveys covering large fractions of the sky, with substantial funding from the NSF. The unique raw data sets delivered by the SDSS are processed through sophisticated software pipelines to produce high-level science-quality images, spectra, and catalogs. All SDSS data products are released publicly through powerful and well-documented interfaces at regular intervals. The latest public SDSS data release (the 13th) encompasses over 125 terabytes of data from 10 coherent survey projects studying the incidence of extrasolar planets, the nature of our Milky Way Galaxy, the evolution of galaxies over cosmic time, and the large-scale cosmology of the universe as a whole. The combination of ambitious surveys, high-quality digital data products, and regular public data releases has led to the SDSS being one of the most fruitful astronomy projects of all time: to date, SDSS data are featured in over 7,500 publications with over 360,000 total citations.

New survey projects—scientific successors to the SDSS—are now bringing this data-intensive mode of astronomy to the NSF-funded National Optical Astronomy Observatory (NOAO). In the present, the Dark Energy Survey (DES) is using a purpose-built wide-field camera on the 4-meter Blanco Telescope at NOAO's Cerro Tololo Inter-American Observatory to conduct one of the widest and deepest imaging surveys of the sky to date. In the near future, the Dark Energy Spectroscopic Instrument (DESI) will install and operate a spectrograph capable of observing 5,000 stars, galaxies, and quasars simultaneously with the 4-meter Mayall Telescope at NOAO's Kitt Peak National Observatory. Both DES and DESI are joint efforts of the NSF and the Department of Energy (DOE), and both projects are motivated primarily by the quest to understand the nature of "dark energy", the mysterious effect that is causing the expansion of our universe to accelerate. The data sets being delivered by these two survey projects also enable numerous astronomy and astrophysics research projects beyond their key dark-energy science goals, and NOAO is focused on developing the institutional capacity to ensure that the full range of these projects can be pursued by the US astronomical community.

The most ambitious data-intensive survey project in the NSF's current ground-based nighttime astronomy portfolio is the Large Synoptic Survey Telescope (LSST). Currently under construction on the summit of Cerro Pachón in Chile with funding from the NSF's Major Research Equipment and Facilities Construction (MREFC) program, the LSST will survey the entire southern-hemisphere sky every three to four days with an 8.4-meter primary mirror feeding a 3.2-gigapixel camera. The LSST's data flow of 20 Terabytes per night will be processed in near-real time to deliver an "alert stream" of approximately 10 million events per night: every detected astronomical object that brightens, dims, explodes, or moves. Over the course of its 10-year survey beginning in the early 2020's, the LSST will accumulate a deep 9-petabyte catalog of tens of billions of objects in the southern sky, based on over 200 petabytes of imaging data. Astronomical community research with LSST will be conducted directly with these data products (both real-time and static). Enabling broad-based access to capabilities for petascale data discovery, exploration, and analysis is therefore essential to delivering on the LSST's scientific potential to illuminate the structural and dynamical nature of our Solar System, Milky Way Galaxy, and Universe.

The LSST alert stream will open a new window onto transient and variable phenomena in our universe, such as exotic supernova explosions in distant galaxies. To understand the physics of these phenomena, astronomers will need to obtain rapid spectroscopic follow-up observations with the largest ground-based telescopes, such as the twin 8-meter facilities of the Gemini Observatory, an international partnership with majority funding from the NSF. Successful execution of these programs will require flexible event-broker systems that enable filtering, down-selection, and distribution of alerts at the rate and volume of LSST, standards and systems for channeling these alerts into properly configured and scheduled follow-up observations, and robust software capabilities for extracting science-ready data products from these observations.

Looking further to the future, NASA's Wide Field Infrared Survey Telescope (WFIRST)—a wide-field cousin of the Hubble Space Telescope (HST)—promises to take data-intensive survey-scale astronomy to space in the mid 2020's. In contrast with HST and the James Webb Space Telescope (JWST), WFIRST will dedicate the majority of its time to systematic imaging and spectroscopic surveys to study the nature of galaxies, dark matter, dark energy, and extrasolar planets. In anticipation of this challenge, the Space Telescope Science Institute (STScI, the science operations center for HST and JWST, and the lead center for WFIRST's Wide Field Imager) is developing

wide-field imaging survey infrastructure using the existing data set from the Pan-STARRS1 survey project. WFIRST and LSST will ultimately provide complementary data sets with significant overlapping coverage of the sky, with the potential for even greater scientific power to be realized through the simultaneous co-analysis of their images.

While the above challenges and opportunities center on dark-sky astronomy, the NSF-funded National Solar Observatory (NSO) and partners are currently building the Daniel K. Inouye Solar Telescope (DKIST), another NSF MREFC project. The 4-meter DKIST primary mirror and supporting adaptive optics system will feed five complex instruments capable of high-resolution and cadence spectral, photometric, and polarimetric observations of the Sun's chromosphere, photosphere, and corona. The observational capabilities of the DKIST will help resolve long-standing mysteries about solar flares and eruptions, the heating of the solar corona, fundamental magnetohydrodynamics, the generation of space weather, and numerous other topics. The DKIST will acquire 150 million images comprising 3.2 PB of data per year, a rate exceeding that of most large scientific facilities and comparable to that of the LSST. Moreover, the big-data challenges are not solely related to volume, as the significant complexity of the facility, its instruments, and the intended observed phenomena underlie major data processing hurdles whose difficulty rival and exceed those in the nighttime astronomy efforts described above. In this data-intensive era, like the nighttime community, the DKIST thus faces significant challenges that must be addressed and overcome in order to realize its scientific potential.

As the managing organization for NOAO, LSST, Gemini, STScI, and NSO, the Association of Universities for Research in Astronomy (AURA) has a uniquely broad perspective on data-intensive research opportunities across astronomy in the present and future, and on the importance of coordinated investment in cyberinfrastructure to ensure that the US astronomical community is positioned to take advantage of these opportunities.

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

The data-intensive astronomy research challenges described above are generically characterized by petascale data sets that are...
Irreplaceable.
Obtained at significant expense.
Complex in their data and metadata structure.
Collected at geographically remote locations.
Processed by sophisticated and computationally intensive pipelines.
Subject to fine-grained proprietary constraints.
Exposed and distributed through multiple interfaces.
Accessed by large science user bases with diverse use cases and analysis requirements.
Scientifically valuable for decades and centuries into the future.

At the most basic level, these data sets pose a significant challenge of data archiving and preservation. Federally funded astronomical centers have a mission to provide long-term stewardship of large, scientifically valuable data sets over very long time horizons. This mission is traditionally accomplished with on-premises storage systems that are maintained and renewed as part of each center's operating program. However, many centers also explore off-premises colocation or storage-as-a-service solutions—often for redundant mirroring and distribution systems—and would likely opt in to these models more fully if the combination of economy, reliability, and accessibility were to become more attractive.

Large astronomical data sets must also be held in close proximity to significant computational processing resources. The demand for this computing capacity can be separated into two main categories: operational pipelining and science-user analysis. Although there are cases where the distinction between these two categories is blurred, they generally pose distinct sets of requirements on the data centers that host active research-oriented astronomical archives.

Operational pipelining is the application of algorithmic software to calibrate, transform, and reduce raw and intermediate astronomical data into images, spectra, data-cubes, catalogs, and metadata that are suitable for high-level science analysis. In the "classical" mode of PI-driven astronomy, with its smaller and more heterogeneous data sets, pipelining has traditionally been executed by PI-led teams on local

hardware using local copies of the raw input data. In the mode of data-intensive survey astronomy, pipelines are designed, developed, maintained, and operated by dedicated teams within the survey projects themselves, relying on expert knowledge of the telescopes, instruments, calibrations, observing plans, data characteristics, and science requirements of the survey.

Operational pipelining for major survey projects therefore requires the hosting data center to provide clustered computing resources that support complex processing workflows involving: (1) coordination with inbound data transfer from remote observatories; (2) on-demand computing capacity for time-critical pipelining of new data; (3) deep and customized software dependency stacks; (4) wrapping and large-scale execution of scientist-contributed code that may not adhere to best coding practices; (5) I/O-intensive processing steps; (6) continuous integration across development, testing, integration, and operations environments and activities; (7) scalable capacity for periodic bulk reprocessing of the entire accumulated survey data set; and (8) integration with survey-specific systems and practices for the management of input, intermediate, and output data. Taken together, these considerations impose both a set of technical requirements, and the necessity for a close working relationship between cyberinfrastructure providers and survey data pipelining teams.

Science user analysis—the other major source of computing demand for large astronomical survey data sets—encompasses querying, exploration, visualization, and post-processing of high-level data products by individuals or teams motivated by specific scientific questions. Again, in the classical mode of astronomy, this work has generally been done with resources and data sets local to the users themselves. In contrast, the large data sets of survey astronomy have driven a trend towards moving science user analysis into close proximity with a central data archive whenever possible, to minimize the need for many large network transfers and expensive local storage and compute systems.

Science user analysis of astronomical survey data is defined by great diversity in both scale and character. Astronomers interested in selecting a small subset of objects from a large survey impose different requirements than those who wish to do an expensive computation across an entire survey data set. Those who interact entirely with object catalogs held in a database likewise impose different requirements than those who wish to analyze data at the image-pixel level. Current approaches focus on supporting science users at astronomical data centers with multiple services, including: (1) provisioning of science user accounts with associated server-side data-access, personal storage, and compute allocations; (2) synchronous and asynchronous queries against large astronomical catalogs, driven either by forms or by free-form SQL-based syntax; (3) capabilities for server-side staging or download of imaging and spectral data subsets corresponding to query results; (4) support for “notebooks” allowing server-side execution of data-analysis scripts in common high-level languages such as Python and R; (5) “containerization” of user-supplied analysis code for execution on computing resources co-located with the data archive; and (6) service-oriented architectures that allow users to interact with the data through either high-level web interfaces or low-level application programming interfaces (APIs). Demand for these science-analysis resources can be highly variable, peaking in particular immediately following a major public release of survey data. These user-oriented considerations in general drive a separate set of technical requirements from those of operational pipelining, and furthermore entail establishing and managing quality-of-service expectations within the science-user community.

The most successful implementations of the above systems will also enable pipeline developers and science users to gain familiarity rapidly and independently. The cyberinfrastructure platforms for these systems must therefore support documentation, standardization, and federation/modularity to the greatest possible extent.

Wider developments in cloud-computing technology hold significant promise for addressing the research challenge from survey-scale astronomy of flexible demand for colocation between large data sets and significant computing resources. The detailed nature of this challenge as described above likely implies that the most promising solutions will involve hybrid models incorporating on-premises, co-location, academic-cloud, and commercial-cloud resources. These hybrid solutions can be enabled by maximizing standardization and interoperability between different cloud-computing models, and by strategic investment by the NSF in cloud-computing infrastructure that is responsive to the requirements of astronomy and other data-intensive research domains.

Finally, the size of present and future astronomical survey data sets will require further advances in high-speed networking infrastructure. Even if all science-user analysis is conducted in colocation with a central archive, transfer of petascale data sets between major archiving, processing, analysis, and distribution centers will necessarily become routine. And, if computing and storage costs continue to decrease to the point that campus- and research group-level centers can host their own petascale archives and analysis clusters at reasonable cost, requests to “just give me all the data” will become more routine, and will be limited by network capacity.

Question 3 Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

Workforce: Astronomy has a well established tradition of recognition and support for professional career paths based on development of hardware (telescopes and associated digital–optical instruments). As the nature of astronomical research evolves to rely increasingly on large surveys and associated data-intensive research modes, the need for stable long-term career paths in scientific software development is increasing as well. The NSF can recognize this importance both through its grant programs and in the scientific software workforce that is employed at NSF-funded centers and facilities.

Sustainability: Major astronomical research facilities (observatories, telescopes, and instrumentation) typically have multi-decade lifecycles. As a consequence, the software systems required for the processing and analysis of their data are also long-lived, often beyond the horizon of active development and support for their core software dependencies. Long-term astronomical software sustainability is therefore a significant challenge, requiring careful consideration of up-front technology choices, ongoing operation and maintenance practices, and the risks and opportunities of software migration and modernization in mature operating facilities.